

Prompt Learning with One-Shot Setting based Feature Space Analysis in Vision-and-Language Models

Yuki Hirohashi^{1,2}Tsubasa Hirakawa²Takayoshi Yamashita²Hironobu Fujiyoshi²¹OMRON Corp.²Chubu University

yuki.hirohashi@omron.com

{takayoshi, fujiyoshi}@isc.chubu.ac.jp

hirakawa@mprg.cs.chubu.ac.jp

Abstract

By using few-shot data and labels, prompt learning obtains optimal prompts that are capable of achieving high performance on downstream tasks. Existing prompt learning methods generate high-quality prompts that are suitable for downstream tasks but tend to perform poorly in scenarios where only very limited data (e.g., one-shot) is available. We address on this challenging one-shot scenario and propose a novel architecture for prompt learning, called Image-Text Feature Alignment Branch (ITFAB). ITFAB aligns text features closer to the centroids of image features and separates text features with different classes to resolve misalignment in the feature space, thereby facilitating the acquisition of high-quality prompts with very limited data. In one-shot setting, our method outperforms the existing CoOp and CoCoOp methods and in some cases even surpasses CoCoOp’s 16-shot performance. Testing on different datasets and domain, show that ITFAB almost matches CoCoOp’s effectiveness. It also works with current prompt learning methods like MapLe and PromptSRC, improving their performance in one-shot setting.

1. Introduction

Vision-and-language models (VLMs) have demonstrated remarkable zero-shot classification capabilities, thus avoiding predefined label spaces and enabling ad-hoc addition of target labels. VLMs are trained from vast numbers of image-text pairs (such as `img` and `alt-text` tags) crawled from the web, with CLIP [27] and ALIGN [14] using 400M and 1B pairs for training respectively.

However, the zero-shot performance deteriorates when samples that are encountered in a downstream task are either absent or very rare in the pretraining dataset, as compared to full-scratch training models [41]. This phenomenon, in which the training distribution is different from

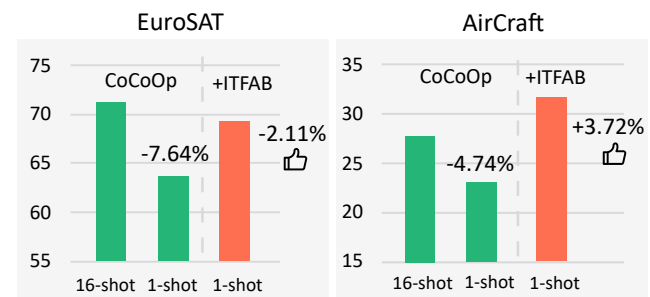


Figure 1: Vanilla CoCoOp performance declines significantly when relying on prompts with only one-shot training. By incorporating our proposed method (ITFAB), models can achieve high performance with only one-shot.

that of the model’s operation, is known as a domain shift. To tackle domain shifts, models are typically adapted to downstream tasks by using few-shot unlabeled or labeled data in the operation environment [9, 34, 39]. On the other hand, VLMs are designed with an enormous number of parameters for training from large-scale training data. As a result, adaptation to downstream tasks is very time-consuming and suffers from the loss of valuable features acquired during pretraining (known as catastrophic forgetting).

One approach to overcome this issue is a prompt engineering, which devises input texts. For instance, in satellite image classification, the prompt “a satellite photo of a `<classname>`” yields 13.3% better performance than the simpler prompt “a photo of a `<classname>`” [41]. However, handcrafting prompts that appropriately describe operation environments is often inefficient. Moreover, such description in practice scenarios may itself be challenging without domain experts.

Prompt learning has been proposed to overcome such limitations of prompt engineering [1, 16, 17, 32, 41, 42]. In prompt learning, prompt tokens are trained using few-shot labeled data for downstream tasks, while the VLM’s image and text encoders are frozen. CoOp [41], a milestone

work in prompt learning, defines prompts as learnable vectors and optimizes them to fit few-shot training samples effectively. Also, to address CoOp’s limited generalization to downstream tasks, CoCoOp [42] has been introduced as a successor model. CoCoOp generates prompts that are conditioned on image features, thus yielding improved generalization capabilities for unseen categories.

Most existing methods assume that at least 16-shots per category are available for prompt learning [16, 17, 42]. In scenarios where data collection is difficult, however, such as medical imaging [24] and industrial inspection [29], it is necessary to adapt to using very limited data (e.g., one-shot data). Figure 1 shows preliminary experiments of 16- and one-shot accuracies within the CoCoOp framework. This analysis reveals a notable decrease in the downstream accuracy for the one-shot setting, as opposed to the conventional 16-shot configuration. Our objective in this paper is to improve downstream accuracy of prompt learning even with very limited data.

To investigate the cause of degradation, we visualized feature spaces of vision-and-language models obtained from prompt learning. The investigation showed that image features form clusters for each category and text features move in the feature space during prompt learning. Also, we found that text features trained by adequate training data (e.g., a 16-shot setting) are located close to the centroids of image features whose category is the same. Conversely, in one-shot prompt learning, these text features demonstrate a limited capacity to move sufficiently from their original positions in the directions of image features’ centroids. We hypothesize that these feature locations are important for high-quality prompts, and propose requirements for the feature space.

Based on above hypothesis, in this paper, we propose Image-Text Feature Alignment Branch (ITFAB), which aims to move text features closer to the centroids of image features, while separating the text features to increase their identification. The proposed branch can encourage exclusion among text features with different labels and inclusion among features with identical labels. This ITFAB mechanism effectively mitigates disparities that arise within the feature space because of limited training data, such as in the case of one-shot learning. Moreover, ITFAB can improve one-shot performance of arbitrary prompt learning methods through integration with existing architectures.

To demonstrate ITFAB’s effectiveness, we conducted one-shot, Base-to-New [42] experiments on 10 different datasets by using CoCoOp architecture. The results confirmed that ITFAB outperformed the conventional methods. On certain datasets, the one-shot accuracy surpassed that of CoCoOp with 16 shots. On the other hand, ITFAB may suffer from overfitting to the source data, because it forces text features to explicitly traverse to the centers of source

data categories. We also evaluated ITFAB’s cross-dataset and domain generalization performance, revealing accuracy comparable to that of CoCoOp. These results show that ITFAB can improve the one-shot accuracy while retaining its generalization capability. Furthermore, we integrated ITFAB into the state-of-the-art (SoTA) prompt learning methods, MapLe [16] and PromptSRC [17], to verify its model-agnostic capacity. Though similar Base-to-New experiments, we confirmed that the one-shot performance was again improved. Our contributions in this paper are twofold:

- We focus on prompt-learning scenarios in which only very limited data is available. From observations of the feature space, we proposed requirements for it that enable high-quality prompts even in a one-shot setting.
- We introduce a branch, ITFAB, that promotes exclusion among features with different labels and inclusion among features with the same label, thus our approach.

2. Related Works

2.1. Vision-and-Language Models for Vision Tasks

Vision-and-Language Models (VLMs) have achieved great success on image recognition fields such as identification, object detection, and semantic segmentation [10, 14, 20, 27, 31, 38]. These models facilitate a convergence of image processing and natural language processing. Specifically, they obtain the correspondence in a feature space of a vast number of pairs of images and text collected from the public web. This approach enables zero-shot prediction without any additional data by matching the positions in the feature space of any image and text [40]. Compared to existing unimodal methods focused only on images, Align [14] and CLIP [27] demonstrate superior zero-shot performance, domain generalization capability, and robustness against adversarial samples.

Although VLMs have outstanding zero-shot performance on various tasks, they are challenging to adapt to downstream tasks without forgetting pretraining knowledge. Recently, two lines of research have been intensively investigated. The first explores transfer learning to downstream tasks, and comprises approaches that emphasize text encoders [16, 17, 41, 42] and those that prioritize image encoders [1, 15]. The second line explores knowledge distillation and aims to improve the model performance on object detection [6, 8], semantic segmentation [5, 21], and so forth. In our work, we investigate methods that focus on applying a text encoder to downstream tasks with only a very limited number of examples, targeting CLIP, one of the most widely used VLMs.

2.2. Prompt Learning

Minor prompt changes, e.g., from ”a photo of $\langle\text{classname}\rangle$ ” to ”a photo of **a** $\langle\text{classname}\rangle$ ” [27], cause VLMs to exhibit

significant accuracy fluctuation on downstream task. The design of appropriate prompts to describe downstream environment is challenging, and domain expert knowledge may be especially needed in industrial and medical domains.

Inspired by “prompt engineering” in natural language processing (NLP), “prompt learning” aims to obtain optimal prompts by applying few-shot samples in a downstream environment. The main approach in prompt learning entails the optimization of prompts representing continuous learnable vectors by using few shot samples. The cross-entropy loss function is often used as the training objective. CoOp [41] was the pioneering attempt to represent prompts as continuous vectors rather than discrete ones. More recently, CoCoOp [42] conditions prompts with image features to enhance generalization for unseen categories. MaPLe [16] innovatively integrates learnable vectors with a text encoder’s intermediate features and subsequently uses the vectors as inputs to an image encoder following a linear transformation. This strategy enables effective multimodal prompt learning. PromptSRC [17] critically addresses the knowledge preservation in pretraining. Specifically, it introduces an advanced prompt-learning model that harmonizes task-agnostic knowledge and task-specific knowledge. Moreover, recent advancements [18] have introduced prompt-learning approaches that circumvent the use of images by leveraging large language models (LLMs). These approaches assume the availability of LLMs or limited testing data, aligning with our research premise. Nonetheless, they lack empirical validation in scenarios that are restricted to one-shot application. Hence, our study primarily concentrates on such scenarios, and propose a novel approach that aims to enhance the one-shot capabilities of current methodologies.

3. Proposed Method

We use the text and image encoders of CLIP as our backbone. Additionally, our prompt learning is based on the most widely used method, which involves vectorizing prompts and learning by minimizing cross-entropy error. This section first introduce these concepts before explaining our proposed method.

3.1. Prompt Learning for CLIP

Prompt learning eliminates the need for handcrafted design of prompt, e.g., “a photo of a”, to match downstream tasks. The earliest work, CoOp [41], defines a prompt as sequence of M continuously differentiable tokens, $[\mathbf{v}_1][\mathbf{v}_2] \dots [\mathbf{v}_M]$. In the case of CLIP-ViT, $[\mathbf{v}_i]$ is 512-dimensional vector. The prompt representing the i th category can be then defined as $\mathbf{t}_i = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M, c_i\}$. Let features from the image encoder and text encoder be denoted by \mathbf{x} and $g(\cdot)$. Then the class probabilities can be expressed by the following formula:

$$p(\hat{y}|\mathbf{x}) = \frac{\exp(\text{sim}(\mathbf{x}, g(\mathbf{t}_y))/\tau)}{\sum_{i=1}^C \exp(\text{sim}(\mathbf{x}, g(\mathbf{t}_i))/\tau)}. \quad (1)$$

Here, $\text{sim}(\cdot, \cdot)$ is a metric that measures similarity in a feature space, with the cosine similarity as a common choice, and τ is the temperature parameter.

CoCoOp [42] conditions prompts with image features to enhance generalization for unseen categories. In practical, image-conditioned prompts $\mathbf{t}_i(\mathbf{x}) = \{\mathbf{v}_1(\mathbf{x}), \mathbf{v}_2(\mathbf{x}), \dots, \mathbf{v}_M(\mathbf{x}), c_i\}$ are formulated by summing meta-tokens π , derived from “meta-net” θ , and the $[\mathbf{v}_i]$. The class probabilities are expressed by

$$p(\hat{y}|\mathbf{x}) = \frac{\exp(\text{sim}(\mathbf{x}, g(\mathbf{t}_y(\mathbf{x}))/\tau)}{\sum_{i=1}^C \exp(\text{sim}(\mathbf{x}, g(\mathbf{t}_i(\mathbf{x}))/\tau)}. \quad (2)$$

Both CoOp and CoCoOp update tokens —CoCoOp additionally adjusts the meta-net parameters— by using the cross-entropy loss from downstream tasks:

$$L_{ce}(y, \hat{y}) = - \sum_{i=1}^C y_i \log(\hat{y}_i). \quad (3)$$

3.2. Feature Space Analysis

In this section, we evaluate the requirements for successful prompt learning by visualizing the feature space and examining the relationship between the locations of image and text features. We compare the feature spaces between one and 16-shot prompt learning in CoOp. Finally, we target the image features $\{\mathbf{x}\}_i^N \in \mathbb{R}^{512}$ obtained from the image encoder, and the text features $\{\mathbf{z}\}_i^K \in \mathbb{R}^{512}$ obtained by passing tokens through the text encoder after prompt learning. Note that N denotes the sample size, while K denotes category size. Note also that, in prompt learning, because each encoder’s weight are frozen, the embedding positions of image features do not change before and after learning; only the embedding positions of the vectorized text features can change.

Figures 2(a) and (b) illustrate a significant disparity in the downstream task performance between one and 16-shot settings. In this study, we seek to understand the reasons for such disparities by analyzing the arrangement of image and text features within the feature space. Using the EuroSAT [11] and OxfordPets [26] datasets as examples, we observed that image features (depicted as dots in the figure) naturally form hyperspherical clusters for each category. This phenomenon is induced by a well-trained CLIP image encoder on a vast amount data, as documented by [30].

In the one-shot feature space, which is highlighted in the areas enclosed by rectangles in Fig.2(a), text features (depicted as stars) either overlapping or are very close to each other. We evaluated the categories and their representative images that are closely embedded in the one-shot scenario.

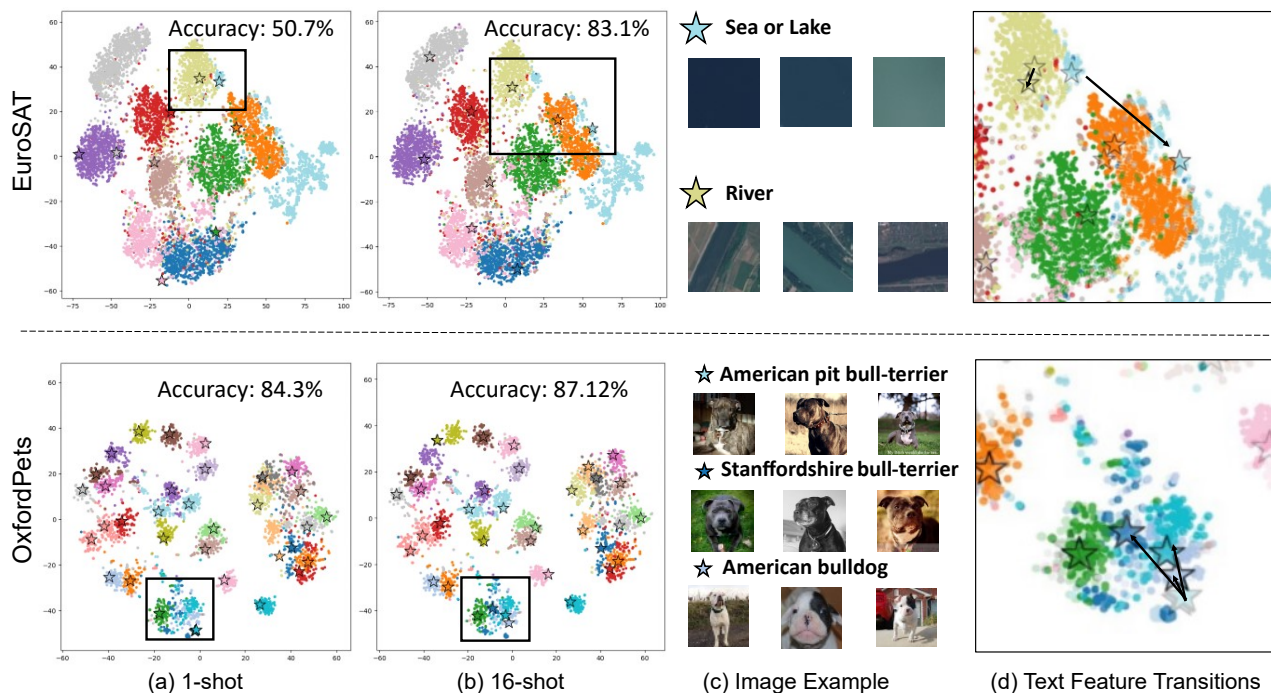


Figure 2: Visualization of the feature space after prompt learning in the (a) one-shot and (b) 16-shot [35]. The stars and dots denote text and image features, respectively. (c) Category names and corresponding sample images indicated by the text features enclosed in rectangles. (d) Zoomed-in view of rectangular region in (a) and (b), showing text features moving from one-shot position to 16-shot position.

Text features with linguistically similar meaning are embedded in close proximity, as shown in Fig.2(c), even when the image characteristics differ significantly. For example, although the image characteristics “Sea or Lake” and “River” differ significantly, both of them have the same meaning “natural water”.

Next, when using a sufficient training data as in the 16-shot setting, text features move closer to the center of image features with the same category, as observed in Fig.2(b). Figure 2(d) provides a zoomed-in view illustrating this shift. CLIP-based models predict categories from similarity between text and image features. Therefore, when text features are embedded close to each other but far from image features with the same category, misrecognition may result in downstream tasks. Below, from these observation, we summarize the requirements for the text and image features in the feature space.

- Text features should be embedded close to the mode of image features with the same category.
- Text features should be embedded far from each other.

3.3. Image-Text Feature Alignment Branch

Our proposed Image-Text Feature Alignment Branch (ITFAB) is based on the above observations. ITFAB promotes prompt updates to simultaneously satisfy the exclusivity of text features with different categories and the inclusivity of

image features with the same category. Our problem setting only allows only one sample per category. With a sufficiently trained image encoder, however, we can assume that the randomly selected single sample is extracted from near the each category’s centroids. By leveraging our hypothesis above, ITFAB can explicitly bring text features closer to the one-shot image features, thereby meeting the requirements outlined in the previous section. Additionally, ITFAB can be integrated with any prompt-learning method, thus improving the performance in one-shot scenarios in the base model. Hence, in this section, we discuss a case where the proposed method is integrated with CoCoOp, a base model that is commonly used in prompt-learning.

Figure 3 illustrates the proposed ITFAB. It is responsible for calculate the loss value, inspired by [23], which augment both the exclusion and inclusion among features from each encoder. For a dataset with K categories, the image encoder’s output features are represented as $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K\}$, with corresponding category $\{y_1, y_2, \dots, y_K\}$. For one-shot learning, the training data volume used in each epoch matches the category count. The text encoder’s output features are denoted as $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K\}$, where these features’ category are equivalent to the category names used within the prompts, denoted as $\{c_i\}_{i=1}^K$. The establishment of $\mathbf{f} = \langle \mathbf{x}, \mathbf{z} \rangle$ and $\mathbf{y} = \{y_1, y_2, \dots, y_K, c_1, c_2, \dots, c_K\}$ enables definition of

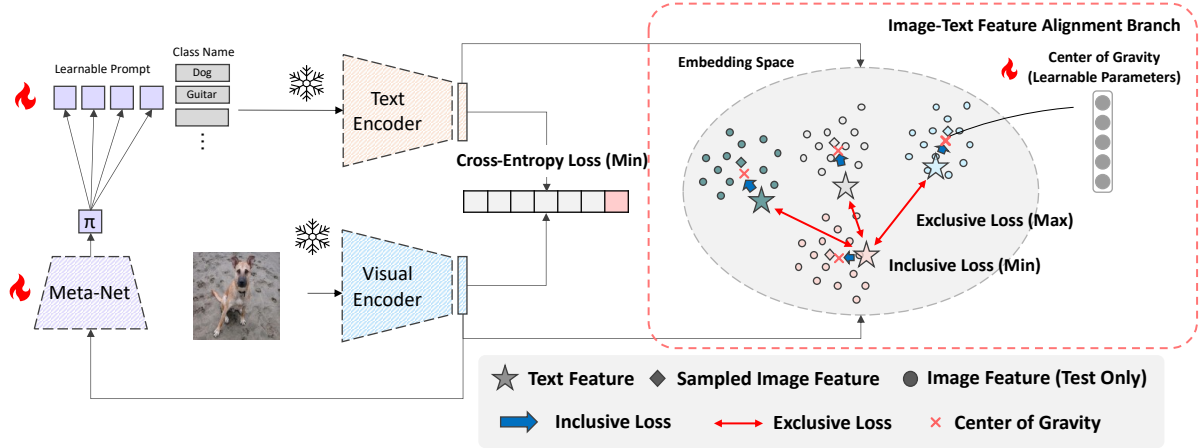


Figure 3: Overview of the model structure integrating Image-Text Feature Alignment Branch (ITFAB) into CoCoOp. Pink dotted rectangle denotes our ITFAB. ITFAB promotes the position alignment of text and image features.

precise inclusion loss \mathcal{L}_{Inc} and exclusion loss \mathcal{L}_{Exc} , as follows:

$$\mathcal{L}_{\text{Inc}} = \sum_i \|\mathbf{f}_i - \mathbf{w}_{y_i}^c\|_2^2 \quad (4)$$

$$\mathcal{L}_{\text{Exc}} = - \sum_i \left(\frac{1}{k-1} \sum_{j \neq y_i} \|\mathbf{f}_i - \mathbf{w}_j^c\|_2^2 + \|\mathbf{w}_{y_i}^c - \mathbf{w}^c\|_2^2 \right) \quad (5)$$

Here, \mathbf{w}^c is a learnable center of gravity, whose position is optimized to the center of each category during prompt learning. \mathcal{L}_{Inc} forces each text feature to move close to the center of gravity, while \mathcal{L}_{Exc} forces each text feature to move far away. These two loss functions are expected to fulfill the above requirements even in one-shot setting.

Finally, as given below, the total loss function $\mathcal{L}_{\text{total}}$ is obtained by adding \mathcal{L}_{Inc} and \mathcal{L}_{Exc} to the base model’s original loss (in this case, CoCoOp’s cross-entropy loss \mathcal{L}_{CE}), which is used for parameter updating. Note that λ_{Inc} and λ_{Exc} are balancing terms.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{Inc}} \mathcal{L}_{\text{Inc}} + \lambda_{\text{Exc}} \mathcal{L}_{\text{Exc}} \quad (6)$$

4. Experiments

In this section, we describe evaluation experiments that were conducted in a benchmark setting to verify the proposed method’s effectiveness. Additionally, we performed ablation study to ensure the validity of the comparisons. Finally, we also assessed the proposed method’s performance when integrated with MaPLE and PromptSRC, the SoTA models in prompt learning.

4.1. Benchmark Setting

Following previous research [42], we conducted the evaluations in a benchmark setting for prompt learning. This benchmark setting includes the Base-to-New, Cross-Dataset Evaluation, and Domain Generalization tasks, which are described below.

Base-to-New

In this task, all categories are divided into two groups: Base and New. The prompt is learned using only the Base categories, and the accuracy on both Base and New test sets is measured using the learned prompt. Previous methods extracted 16 samples per category from the Base, but we tackled the more challenging task of learning the prompt in a one-shot scenario.

Cross-Dataset Evaluation

In this task, prompts trained on ImageNet are evaluated on other datasets. The proposed method may overfit to the source dataset because it explicitly adjust the positions of text features in the feature space. Accordingly, we used this task to verify that our method could achieves performance equivalent to previous methods.

Domain Generalization

In this task, prompts trained on ImageNet are evaluated on ImageNet variants. As mentioned above, our objective was to determine whether generalization capability would be retained.

Datasets

We used 10 classification datasets with the Base-to-New and Cross-Dataset Evaluation tasks. Specifically, we used the general object category datasets ImageNet [4] and Caltech101 [7]; datasets for fine-grained classification such as OxfordPets [26], StanfordCars [19], Flowers102 [25],

Table 1: Comparison with CoOp/CoCoOp on Cross-Dataset Evaluation. CoCoOp with ITFAB (Ours) achieve comparable results with vanilla CoCoOp.

	Source				Target						
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers	Food101	Aircraft	SUN397	EuroSAT	UCF101	Average
CoOp	62.20	83.53	81.30	60.23	58.57	79.50	12.40	52.77	42.13	58.60	59.12
CoCoOp	69.50	93.73	89.47	65.07	70.77	85.97	21.57	65.73	43.63	66.67	67.21
CoCoOp+ITFAB	69.50	93.23	89.50	66.20	70.57	86.10	21.23	66.37	42.17	68.07	67.29

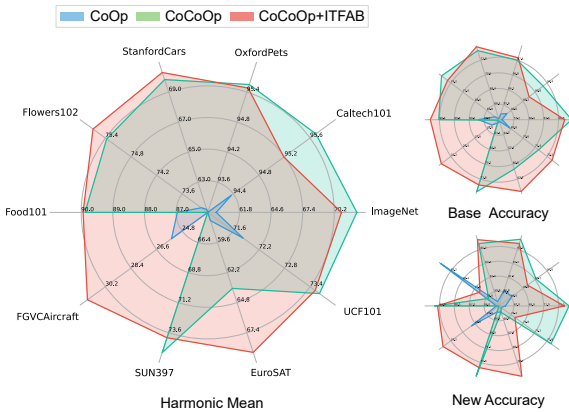


Figure 4: Comparison of ITFAB with CoOp and CoCoOp on Base-to-New. All models were trained with one-shot setting, with one sample chosen from each category.

Food101 [2], and FGVCaircraft [22]; the scene classification dataset SUN397 [37]; the action classification dataset UCF101 [33]; and the satellite image dataset EuroSAT [11]. For the Domain Generalization task, we trained prompts by using ImageNet as the source data and evaluating the accuracy on ImageNet variants; ImageNetV2 [28] (a resampling), ImageNetSketch [36] (a sketch version), ImageNet-A [13] (adversarial images), and ImageNet-R [12] (a wide range of styles, such as sculptures and cartoons).

Previous studies [41, 42] also included the texture dataset DTD [3] in their evaluations. We analyzed each category in DTD in detail and found significant variance in image characteristics even within the same category. For example, the category “matted” includes images with significantly different features, such as lawns, dogs, and perms. Given our problem setting of one-shot learning, such variability within the dataset could cause performance fluctuations, depending on the chosen sample, which would make accurate evaluation difficult. Accordingly, we excluded DTD from the evaluation datasets and used it in an independent evaluation in Section 4.3.

Experimental Setting

We trained prompts with only one-shot, which was randomly selected for each class. We used image and text encoder on a pretrained ViT-B/16 CLIP model. For the benchmark experiments, we integrated ITFAB with the CoCoOp architecture as shown in Fig.3. We trained CoCoOp for 10 epochs in the original manner, followed by five epochs with the proposed losses calculated from ITFAB. The weights λ_{Inc} and λ_{Exc} were set to 0.001 and 0.0001, respectively. Other parameters, such as the batch size and context length, followed by the CoCoOp settings. Each experiment was run three times with different seeds, and the averaged result is shown.

4.2. Main Results

We evaluated ITFAB’s effectiveness by comparing it with CoOp and CoCoOp. First, we verified how our method surpassed existing methods by investigating the Base-to-New task. Then, we demonstrated the ITFAB’s generalization by analyzing Cross-Dataset Evaluation and Domain Generalization task results.

Base-to-New

Figure 4 shows the comparison results for the proposed method with CoOp and CoCoOp across 10 different datasets. These result comprise the accuracies for the Base and New categories and their Harmonic mean. The detailed accuracies for each dataset are given in the Appendix 6. With implicitly operating text features in a feature spaces, ITFAB improve Harmonic Mean on all 10 datasets in comparison with CoCoOp, and obtains an overall gain from 73.39% to 74.55%. The proposed ITFAB increased the average one-shot accuracy across the 10 datasets by +1.16% as compared to vanilla CoCoOp. Regarding the individual accuracy for each dataset, ITFAB demonstrated strong generalization capabilities, notably with improvements of +8.46% and +5.43%, respectively for the Aircraft and EuroSAT.

For other datasets, however, the accuracy was compara-

ble to that of CoCoOp. We attribute the limited accuracy improvement for some datasets to text features’ position in a feature space on initialization. In the case of general categories which extensively include in the CLIP pretraining dataset, initial prompts might be somewhat optimal before prompt learning. As a result, the proposed method’s adjustments to the text features’ position did not yield significant effects.

Cross-Dataset Evaluation & Domain Generalization

For the Cross-Dataset Evaluation and Domain Generalization tasks, we trained prompts on all 1000 categories in ImageNet, and then directly transferred them to the other nine datasets and ImageNet variants. As ITFAB explicitly forces text features’ position closer to the centers of image features, the trained models might have overfit to the source dataset. As summarized in Tables 1 and 2, the proposed method outperformed CoOp and achieved performance comparable with vanilla CoCoOp. These results suggest that the proposed method does not overfit to the source dataset and is capable of solving different datasets and out-of-distribution datasets without performance degradation.

Table 2: Comparison with CoOp/CoCoOp on Domain Generalization. ”-*” depicts suffix of ImageNet variants. CoCoOp with ITFAB (Ours) achieve comparable results with vanilla CoCoOp.

	Source		Target		
	ImageNet	-V2	-S	-A	-R
CoOp	62.20	55.20	40.53	43.30	68.03
CoCoOp	69.50	63.00	48.23	50.30	76.40
CoCoOp+ITFAB	69.50	62.80	48.10	50.00	76.20

4.3. Ablation Study

Epoch Fairness

As mentioned above, we trained our method first with the CoCoOp for 10 epochs and then with proposed losses for an additional five more epochs. To show that our improvements were not just due to the additional training epochs, we also extended CoCoOp’s training to 15 epochs. In Table 3, “CoCoOp” refers to the model trained for the 10 epochs as in the original paper, while “CoCoOp[†]” denotes the model trained for 15 epochs. The values represent the average accuracy across the 10 datasets, similar to the Base-to-New experiment. The results indicate that CoCoOp[†] actually performed worse than original CoCoOp, with a particularly notable performance degradation in the New subset. This result suggests that the increased training cause CoCoOp to overfit to the Base subset, thereby reducing the generalization performance. In contrast, our method main-

tained high performance without losing its generalization capability.

Table 3: Ablation on epoch size. CoCoOp[†] denotes the model trained for 15 epoch.

	Base	New	H
CoCoOp	72.02	72.04	71.97
CoCoOp [†]	72.27	70.88	71.48
CoCoOp+ITFAB	73.07	72.11	72.49

DTD Dataset Evaluation

For the reason explained in Section 4.1, we exclude the DTD dataset from the Base-to-New task and conducted an individual evaluation on it.

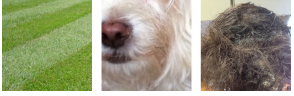





The bottom row of Table 4 summarizes the average performance across all categories in DTD. Overall, the proposed method performed -3.98% worse than CoCoOp did. We measured the accuracy for each category and examined the top three categories with the largest accuracy differences between ITFAB and CoCoOp. The results indicated that the categories, where CoCoOp performed better, exhibited significant variability in image characteristics within the category (for example, the “matted” category includes grass, dogs, and hair). In contrast, the categories, where ITFAB excelled, tended to have relatively less variability within the category.

Our method seeks to improve the one-shot performance by aligning text and image features in the feature space. In cases where image characteristics have large variability within the same category, even if text features are brought close to the image features chosen for the one-shot setting, the probability that other image from the same category appear in similar positions decreases. Without adjusting embedding positions, CoCoOp handles such variability better and outperforms our method in these cases. Conversely, in categories with less variability, whose images are more similar, our method is more effective than CoCoOp, because aligning one image closely aligns others too.

Accuracy Beyond One-Shot Setting

Figure 5 shows the accuracy trends when the number of shots was increased from one to 16 shots. The solid red line represents the harmonic mean for the proposed method, and the solid green line represents that for CoCoOp, with the accuracies for both Base and New indicated by dashed lines. For reference, the results of zero-shot evaluation using CLIP directly is plotted with blue dots. These results demonstrate that the proposed method had the effect of consistently enhancing the backbone model’s performance even when the number of shot was increased.

Table 4: Accuracy and sample images of DTD each category after one-shot prompt learning. The upper half shows categories where CoCoOp’s performance is relatively superior, while the lower half indicates categories where the proposed method excels.

Classname	Accuracy (Ours / CoCoOp / Δ)	Sample images
matted	38.99 / 66.67 -27.78%	
meshed	14.81 / 41.67 -26.85%	
gauzy	20.37 / 46.30 -25.93%	
lacelike	42.59 / 9.26 +33.3%	
potholed	80.56 / 73.15 +7.41%	
spiralled	36.11 / 29.63 +6.48%	
ALL	54.35 / 58.33	-

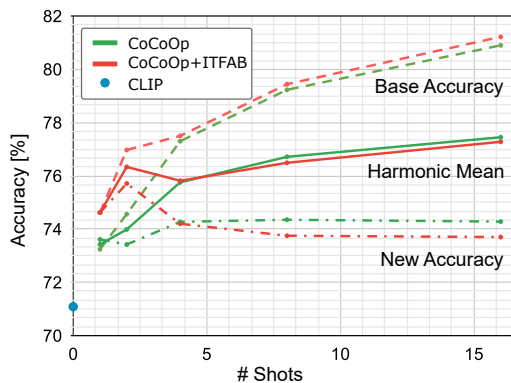


Figure 5: Comparison with CoCoOp from one- to 16-shot on the Base-to-New task across 10 dataset. CLIP’s zero-shot accuracy is also shown for reference.

4.4. Model-Agnostic Study

The proposed ITFAB can be integrated into any prompt learning method and contribute to improve its one-shot performance. To evaluate ITFAB’s model agnosticism, we applied it to other SoTA prompt learning approaches, MaPLE and PromptSRC.

MaPLE: Multi-modal Prompt Learning (MaPLE) [16] learns prompts across both image and text branches, attempting multi-modal optimization. We integrated ITFAB into the MaPLE architecture, by using the features obtained from the image and text encoders.

PromptSRC: Prompting with Self-regulating Constraint (PromptSRC) [17] seeks Pareto optimal solutions between task-specific and task-agnostic knowledge. As with MaPLE, we integrated ITFAB into PromptSRC.

Table 5 lists the average accuracies across 10 datasets in a one-shot setting. All the models with ITFAB showed improved accuracy. MaPLE with ITFAB showed improvement in both Base and New accuracies, with a harmonic mean increase of 0.9%. On the other hand, PromptSRC with ITFAB exhibited a performance decrease of about 1.5% as compared to MaPLE. As PromptSRC seeks the Pareto optimum between pretrained and downstream knowledge, ITFAB might have made it difficult for PromptSRC to converge to the optimal solution because it forces text features to move close to image features.

Table 5: Model agnostic evaluation results. Models integrated with ITFAB improve one-shot accuracy.

	Base	New	H
MapLe	71.88	71.87	71.80
MapLe+ITFAB	72.13	73.38	72.70
PromptSRC	71.37	69.17	70.13
PromptSRC+ITFAB	71.58	69.08	70.19

5. Conclusion

Prompt learning is an effective technique for adapting VLMs like CLIP to downstream tasks using few-shot samples. However, existing prompt learning methods assume that 16 samples per category are available. In this paper, we focused on the more challenging scenario of one-shot prompt learning. First, to investigate the requirements for better performance, we visualized and compared the feature spaces after prompt learning between the one- and 16-shot settings. Leveraging this analysis, we introduced Image-Text Feature Alignment Branch (ITFAB) for alignment of image and text features in a feature space. CoCoOp integrated with ITFAB showed improved one-shot accuracy as compared with vanilla CoCoOp. Furthermore, ITFAB worked with SOTA prompt learning methods like MaPLE and PromptSRC.

References

- [1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. 2022. 1, 2
- [2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 6
- [3] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2013. 6
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5
- [5] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 2
- [6] Yu Du, Fangyun Wei, Zihe Zhang, Miaoqing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 2
- [7] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004. 5
- [8] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, 2022. 2
- [9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. In *JMLR*, 2016. 1
- [10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2
- [11] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS*, 2019. 3, 6
- [12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. 2021. 6
- [13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 6
- [14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. 2021. 1, 2
- [15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2
- [16] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 1, 2, 3, 8
- [17] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, 2023. 1, 2, 3, 8
- [18] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models, 2024. 3
- [19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. 5
- [20] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, 2023. 2
- [21] Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022. 2
- [22] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. 2013. 6
- [23] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *ICCV*, 2019. 4
- [24] Jannatul Nayem, Sayed Sahriar Hasan, Noshin Amina, Bristy Das, Md Shahin Ali, Md Manjurul Ahsan, and Shivakumar Raman. Few shot learning for medical imaging: A comparative analysis of methodologies and formal mathematical framework. 2023. 2
- [25] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008. 5
- [26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 3, 5
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2
- [28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? 2019. 6
- [29] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022. 2
- [30] Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. 2021. 3
- [31] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*, 2022. 2
- [32] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 1

- [33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012. [6](#)
- [34] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. [1](#)
- [35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. In *JMLR*, pages 2579–2605, 2008. [4](#)
- [36] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power. In *PMLR*, 2019. [6](#)
- [37] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. [6](#)
- [38] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Neurips*, 2022. [2](#)
- [39] Qing Yu, Atsushi Hashimoto, and Yoshitaka Ushiku. Noisy universal domain adaptation via divergence optimization for visual recognition. 2023. [1](#)
- [40] Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. 2023. [2](#)
- [41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *IJCV*, page 2337 – 2348. Springer Science and Business Media LLC, 2022. [1](#), [2](#), [3](#), [6](#)
- [42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. [1](#), [2](#), [3](#), [5](#), [6](#)